|  |
| --- |
| Budapesti Műszaki és Gazdaságtudományi Egyetem |
| Diplomamunka |
| [Document subtitle] |

|  |
| --- |
| Bujtor Bálint  [Date] |

Table of Contents

[1. Abstract (3-4 paragraphs): 1](#_Toc56077952)

[1.1. Subject placement, importance of topic 1](#_Toc56077953)

[1.2. Quick summary of what was done and why 1](#_Toc56077954)

[1.3. How it turned out 1](#_Toc56077955)

[2. Introduction (2-3 pages): 2](#_Toc56077956)

[2.1. More detailed introduction to the broad topic 2](#_Toc56077957)

[2.2. Intro into the narrow topic 2](#_Toc56077958)

[2.3. Problem statement (what is important/unsolved that this thesis addresses) 3](#_Toc56077959)

[2.4. Short summary of proposed solution(s) 3](#_Toc56077960)

[2.5. Short summary of results 4](#_Toc56077961)

[2.6. Quick overview of the thesis structure 4](#_Toc56077962)

[3. Related Work (15-20 pages): 5](#_Toc56077963)

[3.1. Give introduction into broad field (e.g: computer vision, deep learning) 5 pages 5](#_Toc56077964)

[3.2. Give introduction into narrow field (e.g: object detection, reinforcement learning) 5 pages 6](#_Toc56077965)

[3.3. Describe chosen algorithm and its alternatives 5 pages 6](#_Toc56077966)

[3.3.1. Make sure you do some pros/cons 6](#_Toc56077967)

[4. Specification (4-5 pages): 6](#_Toc56077968)

[4.1. Describe the actual task in detail 6](#_Toc56077969)

[4.2. Make sure to mention constraints, limitations, requirements 6](#_Toc56077970)

[5. Design (chapters 5-6-7 should be 20-30 pages): 6](#_Toc56077971)

[5.1. Which of the algorithms described in the Related Work chapter we chose and why 6](#_Toc56077972)

[5.2. What frameworks are available and how are they different and why we chose the one we chose 6](#_Toc56077973)

[5.3. Software design (if there is one) 6](#_Toc56077974)

[6. Development: 6](#_Toc56077975)

[6.1. What Programming language/IDE/Environment we used 6](#_Toc56077976)

[6.2. Dataset creation (if there is one) 6](#_Toc56077977)

[6.3. Describe implementation of modules 6](#_Toc56077978)

[6.4. Describe algorithms/neural net structure 6](#_Toc56077979)

[6.5. Describe different attempts 6](#_Toc56077980)

[7. Experimental Results: 6](#_Toc56077981)

[7.1. Describe experiment methodology 6](#_Toc56077982)

[7.2. Describe test cases 6](#_Toc56077983)

[7.3. Describe metrics used for evaluation 6](#_Toc56077984)

[7.4. Describe dataset (if there is one and it wasn't created) 6](#_Toc56077985)

[7.5. Describe test cases 6](#_Toc56077986)

[7.6. Evaluate results 6](#_Toc56077987)

[8. Conclusion (1-2 pages): 6](#_Toc56077988)

[8.1. Give general conclusion (what we did and why is it good) 6](#_Toc56077989)

[8.2. Evaluate results quickly 6](#_Toc56077990)

[8.3. Describe opportunities for further research/improvement 6](#_Toc56077991)

# Abstract (3-4 paragraphs):

Szakdolgozatom témáját a mesterséges intelligencia területéről, azon belül a deep learning témájából írtam. Feladatnak azt választottam, hogy lehetséges-e egy valós környezetben is futtatható objektum detektáló algoritmust mesterségesen szintetizált képeken tanítani, majd pedig a betanított algoritmust valós környezetben sikeresen tesztelni, alkalmazni.

## Subject placement, importance of topic

A mesterséges intelligencia napjaink egyik igen felkapott témája, rengeteg kutatás történik ezen a tématerületen, számos és egyre több szakterületen alkalmaznak sikeresen mesterséges intelligencia alapú algoritmusokat, szolgáltatásokat. Az feladatom is ehhez a népszerű témakörhöz tartozik. A témával foglalkozó kutatások az utóbbi években egyre inkább kezdenek megjelenni a mindennapi életben, gondoljunk az egyre nagyobb autonómiával bíró vagy akár teljesen autonóm járművekre. Ezek mindennapi jelentősége hatalmas. Egy biztonságosan működő autonóm jármű esetén a közúti balesetek száma jelentős mértékben csökkenthető, hiszen a számítógép vezérlése által elimináljuk az emberi mulasztásból, hibából fakadó baleseteket. (ESETLEG VALAMILYEN GRAFIKON, KÖZÚTI BALESETEK SZÁMA, EMBERI MULASZTÁS MÉRTÉKE, MINT OKOZÓ)

## Quick summary of what was done and why

Az én munkám egy mesterséges intelligencia alapú algoritmus megalkotása volt, amely képes valós időben objektumokat detektálni. A mesterséges intelligencián alapuló algoritmusok sikeres betanításának alapja a megfelelő minőségű és pontosságú adatbázis megalkotása. Ebből kifolyólag egy új feladat során ennek az adatbázisnak az előállításával telik el a legtöbb idő. Az én feladatom az adatbázis előállításával eltelő időt eliminálja azáltal, hogy olyan szintetikusan generált képeken tanítja be a mélytanuláson alapuló algoritmust, aminek segítségével az adatbázis generálás automatikusan, emberi felügyelet nélkül történik meg. Ezáltal nagy mértékben rövidíteni lehet egy új mesterséges intelligencia alapú algoritmus kifejlesztését. A feladatom során valós időben kell detektálnom embereket, járműveket, közúti táblákat, illetve jelzőlámpákat.

## How it turned out

# Introduction (2-3 pages):

A következő néhány oldalon egy rövid bevezetőt fogok adni a tématerületről, részletesebben bemutatom a feladatot. Ki fogok térni arra is, hogy pontosan milyen problémára kellett megoldást találnom, valamint arra is, hogy ezt milyen eredménnyel sikerült elvégeznem.

## More detailed introduction to the broad topic

A szakdolgozatom témájának területét a számítógépes látás és a mesterséges intelligencia alapú rendszerek alkotják. A számítógépes látórendszerek olyan algoritmusok segítségével oldanak meg problémákat, amelyeknek bemenete egy adott kép vagy képsorozat, kimenete pedig valamilyen magasabb absztrakciós szintű információ vagy egy módosított kép, amely a felhasználó számára valamilyen szempontból előnyösebb. Utóbbi esetben képfeldolgozásról beszélünk (image processing), előbbi esetben pedig számítógépes látásról (computer vision). A számítógépes látást gyakorta két csoportra szokták osztani, a hagyományos látásra és a gépi tanuláson alapuló algoritmusokra (machine learning). Ezeket a módszereket szokás mesterséges intelligencia alapú (artificial intelligence – AI) algoritmusoknak nevezni.

## Intro into the narrow topic

Az utóbbi években jelentős mértékben nőtt a népszerűsége az úgynevezett mélytanuláson (deep learning) alapuló számítógépes rendszereknek. A mélytanulás a gépi tanulás egyik módszere. A technika lényege, hogy azt tanítjuk meg a számítógépnek, amit mi emberek természetesen csinálunk. Példákon keresztül tanulunk. A mélytanulás olyan technológiák alapja, mint az önvezető autók, hangfelismerő rendszerek (Shazam) vagy akár tőzsdei elemzőprogramok.

A deep learning segítségével olyan komplex problémák megoldása vált lehetségessé, amelyeket a hagyományos számítógépes látáson alapuló programokkal egyáltalán nem, vagy csak óriási erőfeszítések árán sikerült volna megoldani. Az ilyen algoritmusok előnye, hogy segítségükkel olyan problémákat is meg lehet oldani, amelyeknek megoldását magunk sem ismerjük. Ezek a problémák általában magas absztrakciós szintűek. A deep learning alapú algoritmusok hátránya, hogy maga az algoritmus egyfajta feketedobozként funkcionál, nem teljesen tudjuk, hogy pontosan milyen folyamatok játszódnak le benne, hogyan jutott el az algoritmus oda, hogy megfelelő konfidenciával megoldjon egy adott feladatot. A mélytanuló algoritmusok rendszeresen alkalmaznak numerikuson optimalizálást vagy statisztikai módszereket, emiatt nem lehetnek 100%-an pontosak.

A mélytanuló algoritmusokat többféleképpen lehet csoportosítani. Az egyik legelterjedtebb felosztás szerint megkülönböztethetünk felügyelt tanítást, felügyelet nélküli tanulást, valamint megerősítéses tanulást. Mindegyik csoportnak megvannak az előnyei és hátrányai is, éppen ezért más-más területeken terjedtek el sikeresen.

A mélytanulás legegyszerűbb fajtája a felügyelt tanulás, amelynek során minden egyes bemenet mellé tartozik egy elvárt kimenet is. Ilyenkor azt várjuk el, hogy az algoritmus minél nagyobb arányban, minél pontosabban tanulja meg, hogy egy adott bemenethez mi a megfelelő, elvárt kimenet. Ennek a tanulási módszernek azonban számos korlátja, gátja van. A felcímkézett tanító adatbázis előállítása rendkívül költséges, időigényes feladat, minősége pedig nagymértékben meghatározza a tanítás sikerességét is. Másik hátránya pedig az, hogy csak olyan problémákat tudunk ezekkel az algoritmusokkal megoldani, amelyeknek a megoldását mi magunk is ismerjük.

Létezik felügyelet nélküli tanulás is, amelynek során a tanuló algoritmus számára csak bemeneti adatokat szolgáltatunk, elvárt kimeneteket nem adunk meg. Ilyenkor a cél az, hogy az algoritmus valamilyen séma alapján maga alkossa meg a kimenetet, valamilyen kompakt módon. Előnye ennek a módszernek, hogy az adatbázist rendkívül olcsón, akár automatikusan is elő lehet állítani, valamint olyan problémák is megoldhatók vele, amelyeknek megoldását mi sem ismerjük. HÁTRÁNY??

A gépi tanulás harmadik fajtája a megerősítéses tanulás (reinforcement learning). Az előbbi két módszertől ez több dologban is különbözik. A megerősítéses tanítás során az esetek döntő többségében összefüggő döntéseket, döntéssorozatot kell meghozni és a döntések helyességéről nem minden esetben kap visszajelzést. A másik különbség a visszajelzés mibenlétében rejlik. Ez a visszajelzés csak azt mondja meg, hogy mennyire volt jó az adott döntés, azt nem mondja meg, hogy mi lett volna a helyes döntés. A megerősítéses tanulás alkalmazási területére egy remek példák a különféle játékok (sakk, számítógépes játékok) vagy a járműirányítási feladatok.

## Problem statement (what is important/unsolved that this thesis addresses)

A szakdolgozatom témája a felügyelt tanulás köre a mélytanuláson belül. Mint azt az előző hasábokon említettem, a felügyelt tanítás egyik nagy hátránya, hogy a tanító adatbázis előállítása és felcímkézése egy rendkívül költséges, időigényes folyamat. MÉG MIT LEHET??

## Short summary of proposed solution(s)

Az én munkám erre a problémára kínál egy megoldást. A munkám során ezt a tanító adatbázist szintetikus, számítógépesen generált képekből hozom létre, valamint az egyes képek felcímkézését is automatikusan végzem el. A cél egy olyan algoritmus létrehozása, amely képes egy járművön futó beágyazott környezetben, valós időben objektumokat detektálni. A közúti környezetben használt mélytanuló algoritmusokat leginkább olyan problémák megoldására szokták alkalmazni, amelyek valamilyen módon javítják a vezető komfortérzetét, növelik a biztonságot vagy terhet vesznek le a vezetőről. Ilyen probléma lehet az automatikus vészfékezés, ha például egy gyalogos vagy biciklis halad át a jármű előtt. További példa lehet az automatikus távolságkövetés, illetve a tábla- és jelzőlámpa felismerés és ezáltal a közúti szabályok betartása. Éppen ezért munkám során azokat az objektumokat igyekeztem betanítani az algoritmusomnak, amelyek a fent bemutatott problémák szempontjából elengedhetetlenek. A mélytanuló algoritmusomat a következő objektumok felismerésére tanítottam be: gyalogos, jármű közlekedési tábla és jelzőlámpa.

Fontos szempont volt a feladatot és az algoritmust tekintve, hogy valós időben fusson ezáltal a mindennapi környezetben, közúton is használni lehessen. Olyan algoritmust kellett választanom és használnom, ami valós időben dolgozza fel a beérkező adatokat. Enélkül nem lehetne garantálni éles környezetben, hogy megfelelő időn belül képes lefutni és döntést hozni az algoritmus.

Többször említettem azt is, hogy az algoritmusnak a valós életben is meg kell állnia a helyét. Ezért azt is meg fogom vizsgálni a munkám során, hogy a betanított algoritmus, hogyan szuperál valós környezetben készült képeken. Ezért egy olyan adatbázison fogom tesztelni az algoritmusomat, amely ilyen képeket tartalmaz.

## Short summary of results

## Quick overview of the thesis structure

A bevezetés lezárásaként be szeretném összefoglalni, hogy a következő fejezetekben miket fogok bemutatni. Elsőként részletesen be fogok számolni a munkám tématerületéről, ami a mesterséges intelligencia és a számítógépes látás. Be fogok számolni arról is, hogy a mesterséges intelligenciának milyen módszerét használom a munkámhoz. Kitérek ennek a területnek a bemutatására, a jelenlegi kutatásokra, a legújabban használta metódusokra és módszerekre.

Miután részletesen bemutattam a tématerületemet rátérek a saját munkám részletezésére. Elsőként specifikálom a pontos feladatot és bemutatom a tervezés egyes lépéseit. Ezt követően rátérek a fejlesztésre, részletesen bemutatom az egyes lépéseket, az esetlegesen felmerült problémákat és a megoldásukat. Végül pedig bemutatom az eredményeket.

# Related Work (15-20 pages):

Jelen fejezetben részeletesen be fogom mutatni azt a szakterületet és tématerületet, amit a szakdolgozatom és diplomamunkám végzése során megismertem. Elsőként a szélesebb, általánosabb bemutatót fogok adni a témáról, majd rátérek arra tématerületre, ami szorosabban kapcsolódik a szakdolgozatmunkámhoz.

## Give introduction into broad field (e.g: computer vision, deep learning) 5 pages

Mint ahogy azt már említettem az eddigiekben, a szakdolgozatom témájának gerincét a számítógépes látás és a mesterséges intelligencia alkotja. Az én feladatom ennek a két tématerületnek a mezsgyéjén helyezkedik el, ezért mindkét tudományterületet meg kellett ismernem és először ezt a két területet fogom ismertetni.

### Computer vision

Elsőként a számítógépes látás témáját szeretném röviden bemutatni. A számítógépes látás olyan feladatokkal, problémákkal foglalkozik, amelyek során valamilyen vizuális, képi adatból szeretnénk információt kinyerni és ezt felhasználni. Ez az információ sokféle lehet. Egy adott képből kinyerhetünk távolság adatokat, megszámolhatjuk, hogy egy objektumból mennyi van az adott képen, megszámolhatjuk, hogy hányféle objektum van. A lehetőségek gyakorlatilag korlátlanok. A számítógépes látás célja, hogy ezeket az adatokat képekből, képsorozatokból algoritmikus eszközök segítségével kinyerjük egy másik döntéshozó alkalmazás vagy személy számára. A számítógépes látás egyik legnagyobb kihívása az, hogy óriási (akár milliós) adathalmazzal dolgozik és ezeknek az adatoknak gyakorlatilag végtelen féle kombinációja állhat elő. Éppen ezért ezeknek a problémáknak a megoldásához hatékony algoritmusokra van szükség. Sokáig a hatékonyság, illetve a nem megfelelő számítási kapacitás szabott gátat a számítógépes látórendszerek széleskörű elterjedésének.

A számítógépes látás további nehézsége, hogy mi emberek egy látott kép alapján könnyen tudunk információt kinyerni, ugyanakkor az a folyamat, amelynek során ezt az információt kinyerjük a képből részben vagy akár teljesen tudat alatt történik. Emiatt ezeket a képességeket nehéz algoritmikus lépések sorozatává konvertálni. Ezen felül sokszor más érzékszerveinket is felhasználjuk a döntésünk meghozatalához, amelyet egy számítógépes látórendszer nem tud igénybe venni. Ezen problémák miatt a számítógépes látás területén sokszor használunk heurisztikus megoldásokat, gépi tanulásra vagy optimalizálásra épülő megoldásokat.

A számítógépes látó algoritmusokat sokféleképp szokás csoportosítani, mindazonáltal a legelterjedtebb megoldás az eljárások kimenete alapján történik. Eszerint megkülönböztethetünk képfeldolgozó (image processing) valamint a számítógépes látás (computer vision) algoritmusait. Előbbi esetében a kimenet egy olyan módon feldolgozott kép, ami a felhasználó szempontjából előnyösebb tulajdonságokkal rendelkezik. Ezen algoritmusok kimenete általában valamilyen további algoritmus bemenetét szolgáltatják, esetleg az emberi láthatóságot javítják. Utóbbi esetben olyan kimenetet kapunk, amely magasabb absztrakciós szinttel rendelkező információt szolgáltat. Ilyen információ lehet a képen lévő objektumok száma, színe mérete, de bármi egyéb is kinyerhető, amit szeretnénk.

KELL MÉG INFÓ? MIROL?

A számítógépes látáson belül legtöbbször két csoportot lehet megkülönböztetni, attól függően, hogy milyen módszereket használ fel az információ kinyerésére. Megkülönböztethetjük azokat a módszereket, amelyek a gépi tanuláson (machine learning) alapszanak, valamint azokat a módszereket, amelyek nem használnak gépi tanuló algoritmusokat. Ezeket hagyományos látó algoritmusoknak nevezünk. A következőkben a gépi tanuláson alapuló algoritmusokat fogom bemutatni, mert ilyen algoritmusokat alkalmaztam a munkám során én is.

HASZNÁLJAK ANGOL SZAVAKAT VAGY LECSERÉLJEK MINDENT MAGYARRA?

### Machine learning

A gépi tanulás a számítógépes algoritmusoknak olyan vizsgálata, kutatása, amelynek célja az, hogy olyan algoritmusok jöjjenek létre, amelyek képesek a tapasztalás segítségével fejlődni. [1] A gépi tanulást szokás a mesterséges intelligencia egyik részhalmazának is tekinteni. A gépi tanuló algoritmusok egy adott adathalmaz alapján készítik el az adott problémát megoldó modellt. Ezt az adathalmaz tanító adatbázisnak is szokás nevezni. Ezek a modellek olyan bemenetek alapján is tudnak döntést hozni, előrejelezni, amely bemenetre azelőtt nem voltak felkészítve. A gépi tanulást számos területen tudják használni, mint például az e-mailek szűrésénél vagy pedig a számítógépes látás területén is, mint ahogy azt később be is fogom mutatni.

A gépi tanulás során a cél az, hogy anélkül tanítsunk meg egy adott feladatra az algoritmust, hogy az adott feladat megoldását explicit módon megadtuk volna az algoritmusnak. [2] Egyszerűbb feladatok esetén ugyan könnyen meg lehet mondani az algoritmusnak, hogy mi az elvárt kimenet, hogy hogyan kell működnie, ugyanakkor összetettebb feladatok esetében ez nehézkes lehet. Gondoljunk csak arra, amikor egy videósorozatból szeretnénk információt kinyerni. Egy videó másodpercenként több mint húsz képkockát tartalmaz. Ha minden egyes képkockára meg szeretnénk mondani, hogy mi a helyes kimenet, akkor már egy néhány perces videó esetén is rengeteg időt pocsékolunk el. Ehelyett a gépet segítjük, hogy ő maga alkossa meg az algoritmusát az átadott bemenetek alapján. Azt, hogy a gép kialakítsa a saját algoritmusát sokféleképpen el lehet érni. Legtöbbször ezen megközelítések szerint szoktuk csoportosítani a gépi tanuló algoritmusokat.

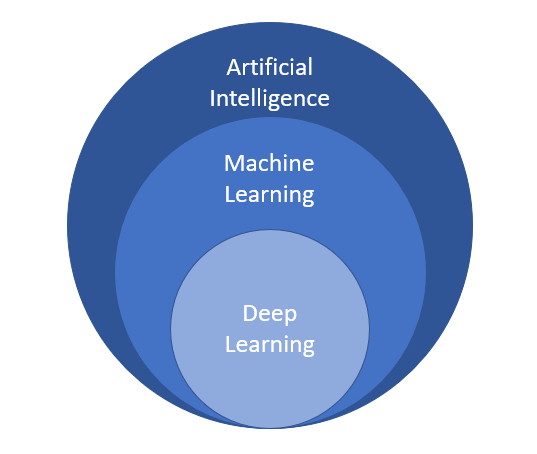
Olyan esetekben például, amikor számos potenciális kimeneti lehetőség létezik, egy lehetséges megoldás az, hogy a bemenetek adott részét érvényesként, a többi részét pedig nem érvényéként címkézzük fel. Erre a módszerre a leggyakrabban hozott példa a számjegyfelismerés. Ebben az esetben a számítógép célja, hogy megtanuljon felismerni kézzel írott számjegyeket. Ilyen kézzel írott számjegyeket tartalmaz az MNIST adatbázis.

A gépi tanulást megközelítő módszereket leginkább három nagy kategóriába szokás csoportosítani, aszerint, hogy milyen természetben található meg jel vagy visszajelzés a tanuló algoritmus számára. Ezeket a kategóriákat már korábban bemutattam, így most csak ismétlés szintjén ejtek szót róluk. Beszélhetünk felügyelt tanulásról, amikor az algoritmus számára egy valaki más által előre felcímkézett adatbázist szolgáltatunk a helyes kimenetekkel. Az cél az, hogy ezen adatbázis segítségével az algoritmus kialakítson egy olyan általános szabályt, amelynek segítségével címkézetlen bemenetekre is nagy százalékban meg tudja adni a kimenetet. A nem felügyelt tanulás során csak egy adatbázist adunk meg az algoritmusnak, elvárt kimenetet nem. Ekkor a célunk az, hogy valamilyen mintát találjon az algoritmus az adatokban. A harmadik kategória a megerősítéses tanulás. Ennek során a program egy dinamikus környezettel érintkezik és a célja az, hogy elérjen valamilyen célt. Minden egyes döntés után kap egy visszajelzést arról, hogy milyen mértékben volt jó a döntése. Ez egyfajta díj, a program pedig ezt a díjat maximalizálja. [3] Természetesen vannak olyan megközelítések is, amelyek nem sorolhatók be a fenti három kategória egyikébe sem vagy egyszerre több kategóriát is alkalmaz a módszer. Ilyen be nem sorolható algoritmus például a dimenzió redukció.

## **History and relationships to other fields**

### Deep learning

As of 2020, deep learning has become the dominant approach for much ongoing work in the field of machine learning. [2]

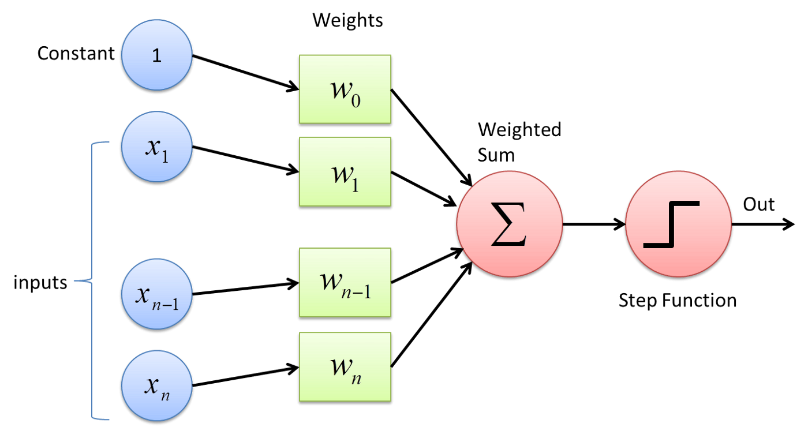
A deep learning fogalmával, tématerületével talán már mindenki találkozott. Napjaink egyik legnépszerűbb és leginkább kutatott témájáról van szó, amelynek alkalmazási lehetőségei különösen szerteágazóak. Népszerűségét és elterjedtségét a rendkívüli rugalmasságának köszönheti, az eljárás kisebb módosításokkal sok helyen jó eredményel meg tudja állni a helyét. Azonban az talán elmondható, hogy legáltalánosabban nagymennyiségű adatok feldolgozásához, értelmezéséhez használják, hogy abból információt nyerjenek ki belőle. Ezek az adatok lehetnek például hangfelvételek, táblázatos adatok, egyszerű adatsorok vagy, mint esetemben, képi adatok, videók.

A deep learning, mint technológia egy nagyobb tématerület részét képezi csak, ezeket a fogalmakat pedig sokszor, sokan szokták összekeverni. A deep learning a mesterséges intelligenciának egyik alfaja, azon belül is a gépi tanulás egyik részhalmaza. Deep learningen belül is megkülönböztetünk különböző válfajokat; így beszélhetünk felügyelt, részben felügyelt, nem felügyelt és megerősítéses deep learning-ről is.

Annak ellenére, hogy a deep learning népszerűsége csak az utóbbi néhány évben kezdett üstökösszerű felívelésbe, maga az eljárás és a téma már régebb óta ismert. Az első működő, felügyelt, több rétegű perceptront 1967-ben publikálta Alexey Ivakhnenko és Valentin Lapa (1)(3). A deep learning fogalmát 1986-ban Rina Dechter (2)(3) vezette be a mesterséges intelligenciával foglalkozó kutatások körébe. A technológia elterjedésének ekkor még elsősorban a rendkívüli adatmennyiség tárolásához és feldolgozásához szükséges tárolási és számítási kapacitás hiánya szabott gátat. Az utóbbi néhány év, évtized robbanásszerű fejlődése az informatikában, a számítási és tárolási kapacitások megnövekedése és az egyre több tárolt adat utat nyitott a deep learning térhódításának.

A deep learning alapjául szolgáló mesterséges neurális hálók ötletét a különböző biológiai rendszerekben megtalálható elosztott kommunikációra és adatfeldolgozásra képes sejtek, pontok inspirálták; gondoljunk csak az emberi agy neuronjaira. Ugyanakkor a szigorúan vett biológia agytól néhány tulajdonságban különbözik is. Míg egy biológiai agy egy élő, változó organizmus, addig egy mesterséges neurális háló statikus és szimbolikus fogalom, amely a működésbeli hasonlóságokra fókuszál.

A deep learning kifejezésében szereplő deep melléknév arra utal, hogy egy ilyen struktúrában több, hasonló réteget használunk fel, ezek a rétegek az úgynevezett perceptronok. Egy perceptron képes eldönteni, hogy a számára megadott számokból álló vektor egy adott osztályhoz tartozik-e vagy sem, amelyet a kimenetén egy bináris számmal jelez. A korai kutatások megmutatták, hogy egy lineáris perceptron nem képes univerzális osztályozásra. Továbbá azt is, ha az ilyen lineáris rétegekből álló hálót ellátjuk egy nemlineáris, aktivációs függvénnyel, akkor az így kapott multilayer perceptron már bármilyen kompakt halmazon folytonos függvényt képes tetszőleges pontossággal megközelíteni, feltéve, ha nem korlátozzuk a rejtett réteg szélességét. Ezt nevezik univerzális approximációnak.

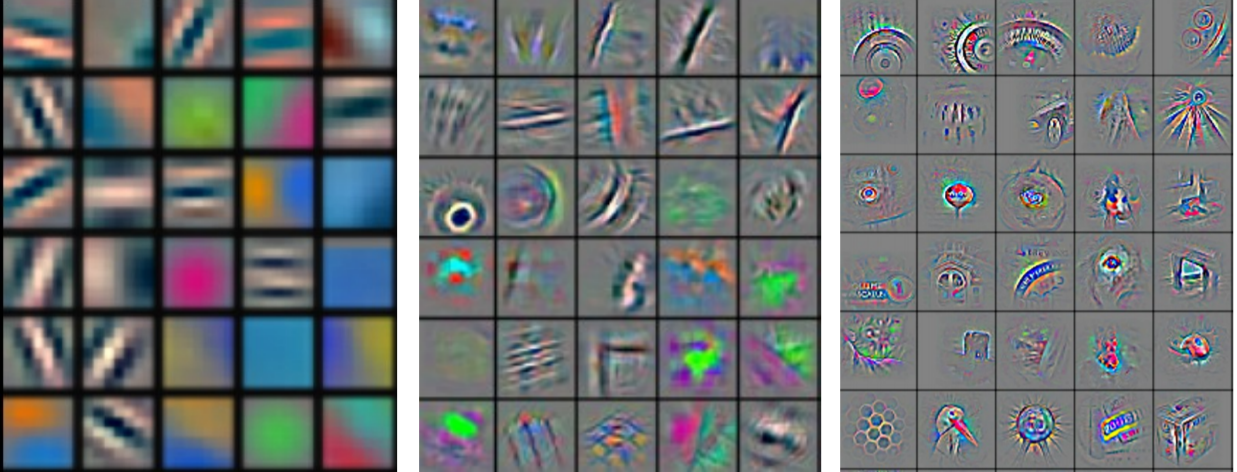


Egy perceptron felépítése, a step function a nemlinearitást hivatott reprezentálni.



Egy Multi-Layer Perceptron felépítése (24)

Az ilyen perceptron rétegekből felépülő neurális háló működését szokás úgy bemutatni, mint hierarchikusan egymásra épülő, egyre komplexebb tulajdonságok kinyerésére képes rétegeket. Például képfeldolgozás esetén az alsóbb rétegek felelnek vonalak, görbék felismeréséért, az azt követő réteg ezekből a formákból alkot alakzatokat (kör, négyzet háromszög). Az ezt követő réteg pedig ezekből a formákból alkotja meg a felismerni kívánt alakzatot, és végül dönti el, hogy az megtalálható-e a bemenetként megadott képen.



A különböző absztrakciós szinteket reprezentáló képek deep learning esetén (balról jobbra egyre növekvő absztrakciós szintek) (4)

Természetesen a mai, modern neurális hálók bőven meghaladták az egyszerű perceptronokból felépülő neurális hálókat, azonban az alapelgondolás gyakorlatilag még mindig ugyanaz. A feladattól függően számos különböző struktúra létezik. A felügyelt tanuláson témakörében képfelismerő algoritmusokhoz előszeretettel használnak CNN-eket (konvolúciós neurális hálókat – Convolutional Neural Network) (5), amelyek az információt konvolúciós neurális rétegek segítségével nyerik ki, a konvolúció műveletének segítségével. Ezekből rendkívül változatos struktúrákat lehet felépíteni, amelyek közül néhánnyal én is megismerkedtem és be is fogok mutatni a későbbiekben. Nem felügyelt tanulással kapcsolatban érdemes még megemlíteni a GAN hálókat (Generative Adversial Networks) (6), amellyel különböző képek eredetiségét lehet vizsgálni, de saját gép által generált képeket is lehet alkotni. A megerősítéses tanulást is számos helyen használják, akár különböző játékokra is.

## Give introduction into narrow field (e.g: object detection, reinforcement learning) 5 pages

A házi feladatom során én is felügyelt tanulást hajtok végre, ezért a következőkben ezt fogom bemutatni. A felügyelt tanulás a deep learning egyik alfaja, amelynek során a DNN-t (Deep Neural Network – mély neurális háló) egy nagy adatbázis segítségével tanítjuk meg adott feladatra. A képi adatbázis minden egyes képéhez tartozik egy címke is, amin azt jelezzük a neurális hálónak, hogy mi az adott bementre elvárt, a háló által adandó kimenet. A neurális háló arra lesz képes, hogy meghatározza, hogy a tanítás során megadott osztályok közül jelen van-e a képen valamelyik.

A feladat az, hogy az elvárt kimenetek függvényében a neurális háló minél nagyobb pontossággal tudja eltalálni azt, hogy az aktuálisan a bemenetére adott kép melyik osztályba tartozik. A háló kezdetben valamilyen véletlen kimenetet ad, hiszen nincs megtanítva a helyes kimenetre. Ezt valamilyen tanító függvény segítségével lehet megtenni. A tanítás során az elvárt kimenet és a háló által adott kimenet között valamilyen különbségképző algoritmussal adjuk meg (például Cross Entropy Loss, MSE Loss) (7), hogy mekkora hibát vét a háló. A hiba alapján egy optimalizáló algoritmus segítségével (például Adam optimization) (8) számítjuk ki, hogy hogyan változtassuk a háló súlyait, hogy minél pontosabb eredményt adjon. Ezt a backpropagation segítségével lehet elérni, amely során a háló neuronjait összekötő kapcsolatok súlyait (és ofszetjeit (bias)) az alapján módosítjuk, hogy mekkora mértékben tér el az elvárt kimenettől. Ezt a gradiens módszerrel tehetjük meg, amely során a parciális deriváltakat számítjuk ki a láncszabály segítségével és annak tükrében módosítjuk a súlyokat és ofszeteket.

A modern felügyelt tanulás alapú neurális hálók (DNN) is ezt a struktúrát követik azonban számos új metódus született annak érdekében, hogy minél jobb eredményt érjenek el. A képfeldolgozás témakörén belül, mint már említettem, a konvolúciós neurális hálók váltak rendkívül elterjedté. A CNN-ek a konvolúció műveletével nyerik ki az információt a feldolgozandó képekből. Az első konvolúciós neurális hálót Kunihiko Fukushima alkotta meg 1980-ban (5). A kép egy adott részletén a konvolúciót elvégezve, abból magasabb szintű információ nyerhető ki, ezzel pedig hierarchikus szinteket lehet felépíteni, hogy valamilyen absztrakt eredményt érhessünk el.



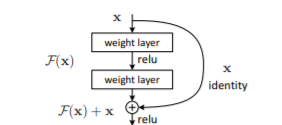
Konvolúciós neurális háló struktúra (24)

Előnye egy lineáris neurális hálóval szemben, hogy az egyes vektorelemek (pixelértékek) egymáshoz képesti pozícióját képes felhasználni, míg ezt egy lineáris réteg nem volt képes kezelni, hiszen az egy egyszerű sorvektorként dolgozta azt fel. További előnye még egyéb algoritmusokhoz képest, hogy az egyes képeket relatíve kicsi előfeldolgozással tudja fogadni. Hátránya, hogy az egyes részletek egymáshoz képesti helyzetét még így sem tudja teljesen jól kezelni (például egy arc esetén a szem, száj, orr egymáshoz képesti helyzetét nem tudja jól felismerni). További hátránya, hogy numerikusan instabil lehet és csak olyan dolgokat képes felismerni, amelyre a tanítás során megtanítottuk (ez minden felügyelt tanulás alapú neurális hálóra igaz). Nem képes tehát saját döntéseket, saját címkéket generálni, olyan információt alkotni, amelyet a megalkotói ne tudtak volna.

A félév során végzett munkám keretében számos modern neurális háló struktúrával is megismerkedtem, néhányat ezek közül ki is próbáltam, mielőtt a végleges algoritmust elkezdtem volna feldolgozni és megérteni. A megismert struktúrák között szerepel a ResNet (9) és az Inception (10)(11), melyek közül a ResNettel foglalkoztam részletesebben.

### ResNet:

Az általános egymás után helyezett konvolúciós rétegekkel ellentétben a ResNet rendelkezik úgynevezett reziduális kapcsolattal, amely egy előre csatoló ágat jelent a struktúrában. Ezzel elérhető, hogy a konvolúció műveletek miatt elvesző információkat visszanyerhessük, így a háló teljesítménye növelhető. Az eddigi neurális hálókkal ellentétben a ResNet képes hasonlóan jó eredményt produkálni sokkal kevésbé mély struktúrával, ezáltal ugyanahhoz a művelethez kevesebb számítási kapacitást igényel. Ugyanakkor képes arra is, hogy egy olyan mély struktúra esetén is megőrizze a numerikus stabilitását, ahol már egy klasszikus, reziduális rétegekkel nem rendelkező háló nem lenne képes, mindezt úgy, hogy még így is kevésbé komplex.



Reziduális építőegység (9)

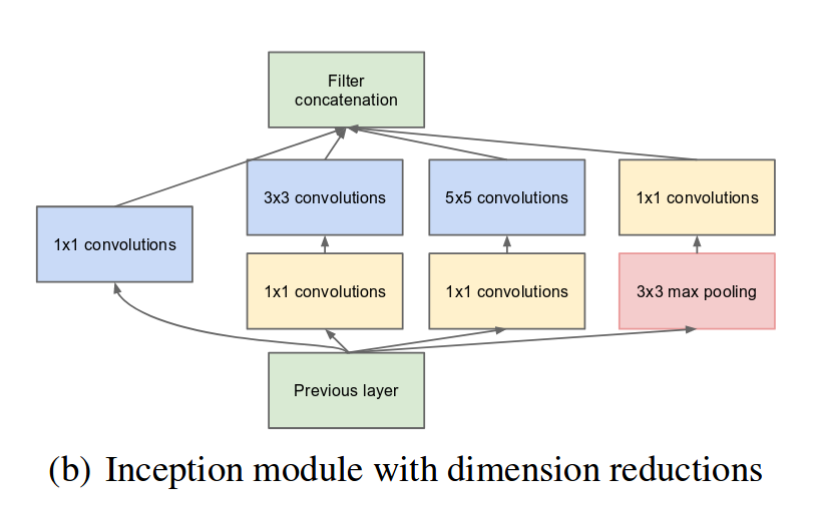
### Inception net:

A másik modern konvolúció neurális háló struktúra, amivel tüzetesebben foglalkoztam az Inception Net (GoogleNet). Egy képen szereplő osztály detektálásakor a számunkra fontos képrészlet rendkívül változatos lehet a méretét tekintve. Éppen ezért a megfelelő kernelméret kiválasztása rendkívül nehéz lehet. Továbbá a rendkívül mély neurális hálók hajlamosabbak az overfitting jelenségére (a tanító adatsoron remek pontosság, de a validációs és tesztadatokon rossz eredmény), valamint a számítási kapacitást tekintve is drágábbnak minősülnek.



A kép különböző részét kitöltő, ámde ugyanazt az osztályt (kutya) ábrázoló képek

Az Inception Net struktúra erre kínál megoldást. Az elgondolás alapja az, hogy egy adott szinten nem egy konvolúciót hajt végre, hanem egyazon szinten több különböző kernel mérettel is elvégzi a konvolúciót (továbbá egy max poolingot is végrehajt), majd az elvégzett konvolúciók kimenetét konkatenálja és ezt adja tovább a következő ilyen rétegnek. Ahhoz, hogy a számítási kapacitást csökkentse, a réteg egy extra 1x1-es konvolúciót is végrehajt a rétegen belül, ezzel limitálva a bemeneti csatornaszámot. Noha ez ellentmondásosnak tűnhet, hiszen pluszműveletet jelent, egy 1x1-es konvolúciót sokkal kevésbé költséges elvégezni, mint egy 5x5-s konvolúciót.

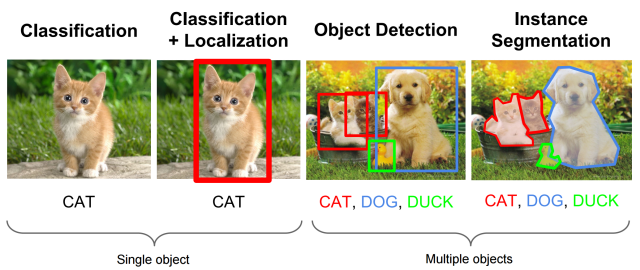


Az inception háló (Inception v1 – GoogleNet) építőegysége dimenzióredukcióval ellátva (11)

A teljes neurális háló ilyen építőegységekből épül fel, természetesen további kiegészítésekkel, amelyekre most nem térek ki, mert nem képezik szerves részét a beszámolómnak. Az struktúrát a későbbiekben tovább is bővítették, finomították, így született meg az Inception v2 és v3.

## Objektum detektálás

Az előző bekezdések során ismertetett neurális hálók objektum osztályozó neurális hálók voltak, ami azt jelenti, hogy rendkívül jó pontossággal képesek megmondani, hogy egy adott osztály szerepel-e a képen vagy sem. Azonban vannak jelentős hátrányaik, amelyek az én témám esetében nagy jelentőséggel bírnak. Nem képesek kezelni azt, ha egy adott képen több osztályból is szerepelnek objektumok, valamint arra sem képesek, hogy a képen belül meghatározzák, hogy hol helyezkednek el ezek az objektumok.



Különböző objektum meghatározási módszerek

A témámhoz tehát egy olyan neurális hálóra struktúrára volt szükségem, amely képes egyszerre több objektumot felismerni és azok helyzetét meghatározni. A félév elején kiválasztott algoritmus a YOLO objektum detektálásra képes struktúra volt, ugyanakkor a félév során más, hasonlóan működő struktúrákkal is megismerkedtem. A következőkben bemutatom az egyes objektum detektáló módszereket, néhány példát, ezután pedig a YOLO architektúrát és működését fogom bemutatni részletesebben.

A különféle objektum detektáló algoritmusokat két csoportba lehet sorolni. Az első csoport a klasszifikáción alapuló algoritmusok. Ezeket az algoritmusokat működésük szempontjából két lépésre lehet bontani. Első lépésként kiválasztanak a képen egy, az algoritmus alapján érdekesnek talált régiót, második lépésként pedig ezeket a régiókat osztályozzák valamilyen CNN segítségével. Ez a módszer nem kifejezetten gyors, hiszen minden kiválasztott régión le kell futtatni az osztályozó algoritmust, ugyanakkor pontosabbak lehetnek, mint a másik alcsoport. Ehhez a kategóriához tartozik például a régió alapú neurális háló (Region-based convolutional neural network (RCNN)) (15) és ennek a továbbfejlesztett verziói, a Fast-RCNN (16), Faster-RCNN (17), valamint a legújabb fejlesztés, a Mask-RCNN (20). Egy másik példa erre a típusra a RetinaNet (18).

A másik csoport a regresszión alapuló algoritmusok. Ezek az algoritmusok ahelyett, hogy kiválasztanák a kép egy adott részét és azon végeznék el az osztályozást, az egész képen próbálják meg előrejelezni az objektumosztályokat és az azt övező bounding box-okat, mindezt egy egyszeri futtatás során. A két leginkább elterjedt algoritmus ebben a kategóriában a YOLO (You Only Look Once) (12)(13) algoritmus, illetve az SSD (Single Shot Multibox Detector) (19). Ezeket az algoritmusokat elsősorban valós idejű alkalmazások során használják, hiszen a pontosságbeli hátrányukat a sokkal gyorsabb futási sebességükkel kompenzálják.

A fenti megfontolások alapján tehát egy regresszión alapuló, valós időben futtatható algoritmusra volt szükségünk, a választásunk pedig a YOLO algoritmusra esett.

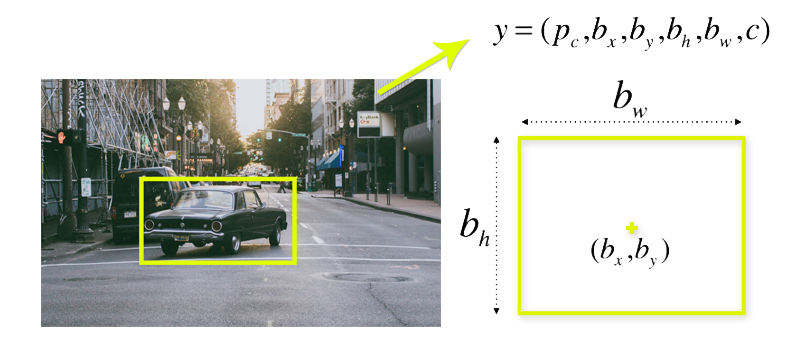
## Describe chosen algorithm and its alternatives 5 pages

A YOLO algoritmus (You Only Look Once) egy hatékony, valós időben futó objektum detektáló algoritmus, amelynek első verzióját Joseph Redmon és társai publikáltak 2015-ben. (12) Az algoritmus egy objektum detektálásra képes metódus, tehát egy képen több osztály több objektumát is képes észlelni, és azok köré egy úgynevezett bounding box-ot (körülvevő keretet/ dobozt) helyezni. Később több, további verzióját, továbbfejlesztését is publikálták az algoritmusnak, a legutóbbi a YOLOv3 (13), amit 2018-ban publikáltak, a következőkben ennek a működését fogom bemutatni.

Mint a legtöbb modern DNN a YOLO is konvolúciós neurális háló (CNN), ezen belül egy teljesen konvolúciós neurális háló (FCN), ami azt jelenti, hogy csak konvolúciós rétegekből épül fel. Összesen 53 konvolúciós réteget tartalmaz, skálázott (strided) rétegekkel és előrecsatoló (reziduális) kapcsolatokkal.

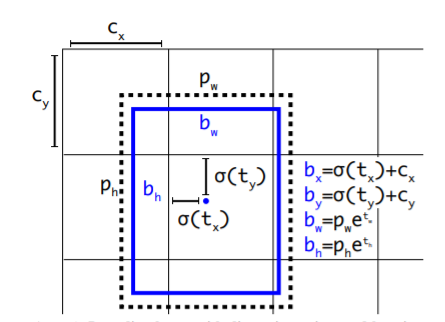
Az eddig megismert struktúrák esetén, a konvolúciós rétegek által megtanult kimeneti valószínűség vektort valamilyen lineáris réteg adja ki. A YOLO esetén ez nem így van, a detektálást egy 1x1-es konvolúciós réteg végzi el, tehát ugyanakkora lesz a kimeneti „kép”, mint az azt megelőző réteg. Ennek a kimeneti képnek a csatornaszáma összesen (B \* (5 + C)), ahol B az adott cella által meghatározott bounding box-ok száma (YOLOv3 esetén ez a szám 3), C pedig az adatbázisban szereplő osztályok száma. Több bounding box-ot azért használunk egy cellán belül, hogy több objektumot is észlelni tudjunk, amelyek ugyanabba a cellába esnek.

A végső célunk az, hogy egy adott osztályba tartozó objektumot észleljünk, továbbá azt is meghatározzuk, hogy az hol helyezkedik el a képen. Ezt a már említett bounding box-ok segítségével tehetjük meg, amit négy paraméter segítségével lehet leírni. Ezek a bounding box szélessége (bw), magassága (bh), valamint a bounding box középpontjának koordinátái (bx, by). Az algoritmus ezen felül kimenetként megadja, hogy mekkora valószínűséggel van jelen egy objektum az adott bounding box-ban, illetve, hogy az mekkora valószínűséggel egy adott osztály tagja.



A YOLO algoritmus bounding boxának paraméterei, valamint a kimeneti vektor elemei (13)

Az algoritmus a képet azonos méretű cellákra osztja (a képtől és a leskálázástól függően), és minden egyes cellára meghatározza a fent említett kimenetet. Azt várjuk el, hogy minden egyes cella detektáljon egy ilyen objektumot abban az esetben, ha a bounding box-ának középpontja az adott cellába esik. Annak elkerülésének érdekében, hogy az adott koordináta ne lóghasson ki a cellából, a bounding box-ot meghatározó paramétereken egy transzformációt hajtanak végre.

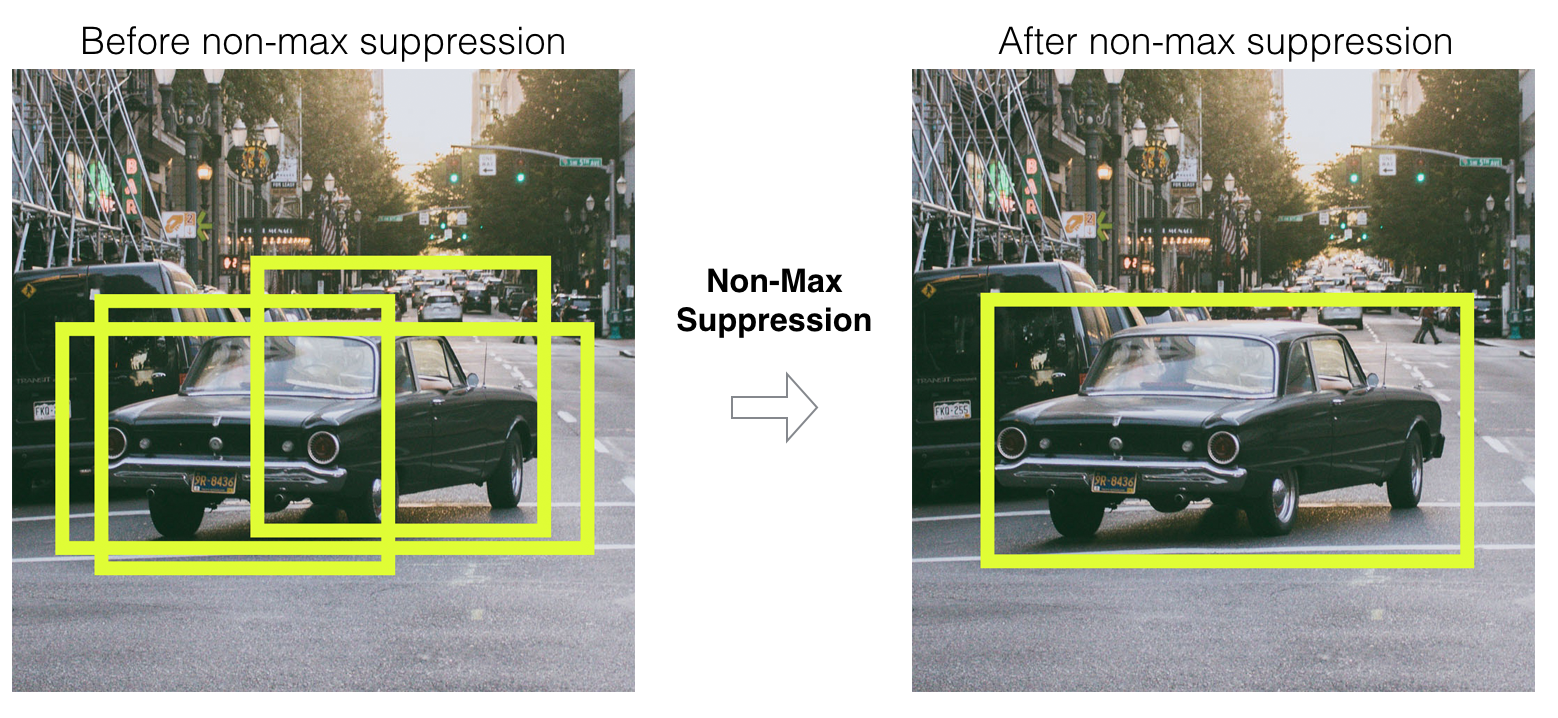


A YOLO algoritmus bounding box leképezés transzformációja (12)

A fenti kép alapján tehát a háló a tx, ty, tw és th paramétereket határozza meg és a fent látható leképezés segítségével kapjuk meg a végleges, fent részletezett paramétereket.

A YOLOv3 algoritmus 3 különböző méretű képen végzi el a detekciókat. Az algoritmus futása során a különböző szintű konvolúciók elvgzése során teszi ezt meg. Nézzük a következő példát, ahol a bemeneti kép 416x416 pixeles, és a három leskálázó réteg méretei: 32,16 és 8. Ez azt jelenti, hogy a kimeneten megkapott képek cellaszámai: 13x13, 26x26 és 52x52. Az algoritmus minden egyes szinten 3 bounding boxot határoz meg minden egyes cellához, ami azt jelenti, hogy összesen ((52 x 52) + (26 x 26) + (13 x 13)) x 3 = 10647 bounding box-ot határoz meg.

Ez a szám rendkívül nagy, természetes nem lesz ennyi objektum a képen. Ezt a számot az algoritmus két lépésben csökkenti le. Elsőként egy bizonyos küszöbszinti alatti valószínűséggel rendelkező detekciókat törli, majd ezt követően Non-Maximum Suppression-t (NMS) hajt végre és így kapjuk meg a végleges, az adott objektumhoz tartozó detekciót.



NMS a YOLO algoritmus során (13)

### Make sure you do some pros/cons

# Specification (4-5 pages):

## Describe the actual task in detail

## Make sure to mention constraints, limitations, requirements

# Design (chapters 5-6-7 should be 20-30 pages):

## Which of the algorithms described in the Related Work chapter we chose and why

## What frameworks are available and how are they different and why we chose the one we chose

## Software design (if there is one)

# Development:

## What Programming language/IDE/Environment we used

## Dataset creation (if there is one)

## Describe implementation of modules

## Describe algorithms/neural net structure

## Describe different attempts

# Experimental Results:

## Describe experiment methodology

## Describe test cases

## Describe metrics used for evaluation

## Describe dataset (if there is one and it wasn't created)

## Describe test cases

## Evaluate results

# Conclusion (1-2 pages):

## Give general conclusion (what we did and why is it good)

## Evaluate results quickly

## Describe opportunities for further research/improvement

# Források, hivatkozások

1. Tom Mitchell: Machine Learning (1997): <http://www.cs.cmu.edu/~tom/mlbook.html>
2. Ethem Alpaydin (2020). Introduction to Machine Learning (Fourth ed.).
3. Bishop, C. M. (2006), Pattern Recognition and Machine Learning, Springer, ISBN 978-0-387-31073-2
4. (1): A. G. Ivakhnenko, Valentin Grigorʹevich Lapa: Cybernetics and forecasting techniques (1967) <https://books.google.hu/books?id=rGFgAAAAMAAJ&redir_esc=y>
5. (2): Rina Dechter: Learning While Searching in Constraint-Satisfaction-Problems (1986): <https://www.researchgate.net/publication/221605378_Learning_While_Searching_in_Constraint-Satisfaction-Problems>
6. (3): Juergen Schmidhuber: Deep Learning (2015): <http://www.scholarpedia.org/article/Deep_Learning>
7. (4): Matthew D Zeiler, Rob Fergus: Visualizing and Understanding Convolutional Networks (2013): <https://arxiv.org/abs/1311.2901>
8. (5): Kunihiko Fukushima: Neocognitron: A Self-organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position (1980): <https://www.cs.princeton.edu/courses/archive/spr08/cos598B/Readings/Fukushima1980.pdf>
9. (6): Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, Yoshua Bengio: Generative Adversarial Nets (2014): <https://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf>
10. (7): Cross Entropy Loss, MSE Loss: <https://pytorch.org/docs/stable/nn.html#loss-functions>
11. (8): Diederik P. Kingma, Jimmy Ba: Adam: A Method for Stochastic Optimization (2017): <https://arxiv.org/abs/1412.6980>
12. (9): Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun: Deep Residual Learning for Image Recognition (2015): <https://arxiv.org/pdf/1512.03385.pdf>
13. (10): Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, Andrew Rabinovich: Going deeper with convolutions (2014): <https://arxiv.org/pdf/1409.4842v1.pdf>
14. (11): Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, Zbigniew Wojna: Rethinking the Inception Architecture for Computer Vision (2015): <https://arxiv.org/pdf/1512.00567v3.pdf>
15. (12): Joseph Redmon, Santosh Divvala, Ross Girshick, Ali Farhadi: You Only Look Once: Unified, Real-Time Object Detection (2015): <https://arxiv.org/pdf/1506.02640.pdf>
16. (13): Joseph Redmon, Ali Farhadi: YOLOv3: An Incremental Improvement (2018): <https://pjreddie.com/media/files/papers/YOLOv3.pdf>
17. (14): <https://github.com/pytorch/vision/blob/master/torchvision/models/resnet.py>
18. (15): Ross Girshick, Jeff Donahue, Trevor Darrell, Jitendra Malik: Rich feature hierarchies for accurate object detection and semantic segmentation (2014): <https://arxiv.org/pdf/1311.2524.pdf>
19. (16): Ross Girshick: Fast R-CNN (2015): <https://arxiv.org/pdf/1504.08083.pdf>
20. (17): Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun: Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks (2016): <https://arxiv.org/pdf/1506.01497.pdf>
21. (18): Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, Piotr Dollár: Focal Loss for Dense Object Detection (2018): <https://arxiv.org/abs/1708.02002>
22. (19): Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, Alexander C. Berg: SSD: Single Shot MultiBox Detector (2015): <https://arxiv.org/abs/1512.02325>
23. (20): Kaiming He, Georgia Gkioxari, Piotr Dollár, Ross Girshick: Mask R-CNN (2017): <https://arxiv.org/abs/1703.06870>
24. (21): Trophy Flux Buggy: <https://www.hpiracing.com/en/kit/107016>
25. (22): Intel RealSense Depth Camera D435i: <https://www.intelrealsense.com/depth-camera-d435i/>
26. (23): Intel® RealSense™ Product Family D400 Series Datasheet: <https://www.intelrealsense.com/wp-content/uploads/2020/05/Intel-RealSense-D400-Series-Datasheet-May-2020.pdf>
27. (24): <https://cs231n.github.io/convolutional-networks/>

Chapters 4-5 can be merged into a "Specification and Design" in case of an algorithm-oriented thesis.

Chapters 5-6 can be merged into a "Design and Development" in case of a hardware-oriented thesis.

**Do:**

* **Always** number your tables, figures and equations.
* **Always**refer to your figures from the text. This tells the reader when to look at them (they might be on the next or the previous page).
* **Always** explain colors and other markings in figures. Preferably in the caption.
* When using abbreviations, expand them for the first time.
* At the start of every chapter, write a paragraph listing what you will talk about in the chapter. This helps the reader to see the underlying structure.

**Don't:**

* **Don't ever**, ***ever***cite Wikipedia. Wikipedia is really good, but it is a collection of references. All it takes is to scroll down to the bottom of the page to find the sources it is citing.
* **Don't** cite lecture slides or notes. Find the paper detailing the algorithm that you learned about in the lecture.
* **Don't** insert equations as images. Sure, it is easy to cut them from the slides using Snipping Tool, but there is a special place in hell for people who do this.
* **Don't** put code, or explain code in the report/thesis. Unless the specific code that you wrote is absolutely brilliant and essential to the value of your thesis. If you want your thesis to be reproducible, put your code on GitHub.